NPS-OR-04-005

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

The Analysis of Shooting Problems via Generalised Bandits

by

Kevin D. Glazebrook
Helen M. Mitchell
Donald P. Gaver
Patricia A. Jacobs

June 2004

Prepared for: Cebrowski Institute and ONR Global

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93943-5000

RDML Patrick W. Dunne, USN
Superintendent

Richard Elster
Provost

This report was prepared for the Cebrowski Institute, Naval Postgraduate School, Monterey, CA 93943 and ONR Global (U.S. Office of Naval Research International Field Office), ONRIFO, PSC 802, Box 39, FPO-AE 09499-0039 USA and funded by the Cebrowski Institute.

This report was prepared by:

KEVIN D. GLAZEBROOK
Professor of Management Science

HELEN M. MITCHELL
School of Mathematics and Statistics
Newcastle University

DONALD P. GAVER
Distinguished Professor of
Operations Research

PATRICIA A. JACOBS
Professor of Operations Research

Reviewed by:

LYN R. WHITAKER
Associate Chairman for Research
Department of Operations Research

Released by:

JAMES N. EAGLE
Chairman
Department of Operations Research

LEONARD A. FERRARI, Ph.D.
Associate Provost and Dean of Research

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.			
1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2004	**3. REPORT TYPE AND DATES COVERED** Technical Report	
4. TITLE AND SUBTITLE: The Analysis of Shooting Problems via Generalised Bandits			**5. FUNDING NUMBERS** BB3CD
6. AUTHOR(S) Kevin D. Glazebrook, H. M. Mitchell, Donald P. Gaver, Patricia A. Jacobs			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			**8. PERFORMING ORGANIZATION REPORT NUMBER** NPS-OR-04-005
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Cebrowski Institute, Naval Postgraduate School, Monterey, CA 93943 and ONR Global (U.S. Office of Naval Research International Field Office), ONRIFO, PSC 802, Box 39, FPO-AE 09499-0039 USA			**10. SPONSORING / MONITORING AGENCY REPORT NUMBER** N/A
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			**12b. DISTRIBUTION CODE**

13. ABSTRACT ***(maximum 200 words)***

A single Red wishes to shoot at a collection of Blue targets in order to maximize some measure of return obtained from Blues killed before Red's own demise. While the class of decision processes called multi-armed bandits has been previously deployed to develop optimal policies for Red, we argue the importance of a little known, but more general class of bandit processes introduced by Nash (1980). In particular, the deployment of this class of processes will enable Red to take account in a natural way of the relative threats posed to his own survival in taking targeting actions. We develop optimal shooting policies for Red in the context of a range of models, which are of independent interest. The paper concludes with a numerical study.

14. SUBJECT TERMS multi-armed bandits, Gittins Indices, suppression of enemy air defense			**15. NUMBER OF PAGES** 26
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UL

The analysis of shooting problems via generalised bandits

K.D. Glazebrook, Management School,
Edinburgh University, UK,

H.M. Mitchell, School of Mathematics and Statistics,
Newcastle University, UK,

D.P. Gaver, Department of Operations Research,
Naval Postgraduate School, Monterey, USA,

P.A. Jacobs, Department of Operations Research,
Naval Postgraduate School, Monterey, USA.

May 14, 2004

Abstract

A single Red wishes to shoot at a collection of Blue targets in order to maximise some measure of return obtained from Blues killed before Red's own demise. While the class of decision processes called multi-armed bandits has been previously deployed to develop optimal policies for Red, we argue the importance of a little known, but more general class of bandit processes introduced by Nash (1980). In particular, the deployment of this class of processes will enable Red to take account in a natural way of the relative threats posed to his own survival in taking targetting actions. We develop optimal shooting policies for Red in the context of a range of models which are of independent interest. The paper concludes with a numerical study.

1 Introduction

A multi-armed bandit problem arises when a single key resource is available for allocation to a fixed collection of projects or bandits. These projects evolve stochastically while in receipt of service (i.e. while the resource is allocated to them) and earn state dependent returns as they do so, but remain fixed (and earn nothing) otherwise. Gittins and Jones (1974) elucidated the optimality of *index policies* for certain classes of multi-armed bandit problems. Such policies attach a calibrating *index* to each project, a function of that project's state, and choose at each decision epoch to allocate resource to whichever project has the largest associated index. See also Gittins (1989). An extensive literature exists outlining a range of extensions and developments of Gittins' classical work while various schemes for index computation have been proposed. See, for example, Whittle (1980), Weber (1992), Katehakis and Veinott (1987) and Bertsimas and Niño-Mora (1996).

Recently, Glazebrook and Washburn (2004) have discussed the utilisation of the multi-armed bandit framework and the associated index policies to develop optimal shooting policies. Here the "key resource" is a single shooter (Red) and the "projects" form a fixed collection of targets (Blue). Red's goal is to so target the Blues as to maximise the expected number (or value) of kills achieved. Manor and Kress (1997) had previously utilised the theory of multi-armed bandits to analyse a shooting problem in which Red receives incomplete information regarding the outcome of successive shots. If a shot is unsuccessful (the Blue target is not killed) then Red receives no feedback, while if the target *is* killed, that fact is confirmed to Red with probability less than one. Manor and Kress (1997) demonstrate the optimality of a form of index policy (the greedy shooting policy) for their setup. Barkdoll et al. (2002) develop index policies for Red in situations in which, not only must he choose which Blue to target, but also how to operate engagement radar in support of each shot.

In a little known early development of Gittins' work, Nash (1980) elucidated the optimality of index policies for a class of *generalised bandits* in which a form of reward dependence is induced between the constituent projects or bandits via a multiplicatively separable structure. Subsequent theoretical developments of Nash's work include those of Fay and Walrand (1991) and Crosbie and Glazebrook (2000 a,b). A general methodology for the computation of Nash's indices may be found in Glazebrook and Greatrix (1995). The prime aim of the current paper is to argue the importance of this class of generalised bandits and their associated index policies for the analysis of shooting problems. As we shall see, they are especially effective in situations in which Red's engagement with the enemy puts him in danger and that the level of danger may differ according to which Blue he targets. In such situations, Red's objective becomes the maximisation of the expected number (or value) of Blues killed before he himself is destroyed. Now Red's shooting policy must balance the returns obtainable from his options against the respective dangers they pose. The index policies we develop elucidate how this balance should be struck.

Consider, for example, a military scenario discussed by Barkdoll et al. (2002) which is asymmetric between the enemy forces. Blue has established air superiority in some region and Red is a surface-to-air missile (SAM) seeking to disrupt Blue's air campaign. The Joint Chiefs of Staff use the terms "reactive" or "opportune" suppression of enemy air defences (SEAD). In Barkdoll et al. (2002) every Red shot at Blue exposed him to danger from a stand-off Blue shooter. Moreover, in such a situation the level of danger to the Red SAM may vary according to the Blue targets he chooses. For example, shooting at longer range puts Red at greater risk to anti-radiation missile (ARM) attack from Blue since the SAM will need to radiate longer to guide the missile to its target. We introduce models and analyses appropriate for situations in which Red's optimal shooting policy needs to take account of such risks to himself.

The paper is structured as follows. Section 2 presents a general class of shooting problems in the form of *generalised bandit* problems. We also describe the nature of the optimising *index policies*. Each of Sections 3-5 feature a particular model along with details of the corresponding optimal shooting policy. Each of these is of independent interest. Model 1 (in Section 3) is a Bayesian model in which Red is able to learn about the (true) identity of the Blues he faces as the engagement proceeds. Model 2 (in Section 4) allows for partial/cumulative damage to each target, while Model 3 (in Section 5) extends Model 1 in allowing Red to supplement the information he has about the Blues he faces by "looking" (imperfectly) at the most recently targetted Blue after each shot. In Section 6, we exemplify the performance of the index policies developed in a numerical study. We conclude in Section

7 with a brief discussion of possible extensions and of issues faced by the Blue force.

2 A General Model

A shooter Red has to plan a series of engagements with N Blues. A single engagement *must* include a shot by Red at a targetted Blue and may expose Red to the possibility of being killed himself. An engagement may also incorporate a look by Red to gain information on the state of the Blue targetted after he has delivered his shot. Red is assumed to have an infinite supply of bullets. His decision problem concerns the choice of which Blue to engage next on the basis of his observational history of past engagements to date. Red's goal is to maximise his expected return from engagements until he himself is destroyed. Red's decision problem is modelled as a Markov decision problem $\{(\Omega_j, \omega_j, P_j, R_j, Q_j, \beta),\ 1 \leq j \leq N\}$ as follows:

(i) $\mathbf{X}(t) = \{X_1(t), X_2(t), \ldots, X_N(t)\}$ denotes the state of the system at time $t = 0, 1, 2, \ldots$ (i.e. before the $(t+1)^{\text{st}}$ engagement) and $X_j(t)$ is the state of Blue j. We require that $X_j(t) \in \Omega_j \cup \{\omega_j\}$, where Ω_j is the space of possible descriptors of Red's knowledge of Blues j's status, while $X_j(t) = \omega_j$ indicates that by time t, Red has been killed during an engagement in which he shot at Blue j;

(ii) At each $t = 0, 1, 2, \ldots$, if Red is still alive he must choose one of the actions $a_1, a_2, \ldots, a_N$. Choice of a_j means that Red's $(t+1)^{\text{st}}$ engagement will target Blue j;

(iii) If action a_j is chosen at t then Red observes a Markovian change of Blue's state $X_j(t) \rightarrow X_j(t+1)$. We write

$$P_j(x, y) = P\{X_j(t+1) = y | X_j(t) = x, a_j\},\ x, y \in \Omega_j \cup \{\omega_j\}.$$

Note that Ω_j may contain a state $\overline{\omega}_j$ indicating that Blue is dead and that a still alive Red knows this. In such cases, both $\overline{\omega}_j$ and ω_j are absorbing states under Markovian law P_j. Note that when action a_j is chosen at t then $X_k(t) = X_k(t+1),\ k \neq j$;

In order to write down expressions for expected rewards, we shall suppose that an infinite string of members of $\{a_1, a_2, \ldots, a_N\}$ are chosen and consequential system state changes (as in (iii)) observed but that rewards can only be collected *while Red is still alive.* To this end, we introduce bounded functions R_j, Q_j and $\widetilde{R}_j$, all from $\Omega_j \cup \{\omega_j\}$ to $\mathbb{R}^+$. The quantity $R_j(x)$ is the expected return secured when action a_j is taken at t and $X_j(t) = x$. Function Q_j is an indicator such that

$$Q_j(x) = \begin{cases} 1, & x \in \Omega_j \\ 0, & x = \omega_j; \end{cases}$$

and $\widetilde{R}_j$ is the product $R_j Q_j$.

Should action a_j be taken at t, the return generated by the ensuing engagement is written

$$\beta^t R_j\{X_j(t)\} \prod_{k=1}^{N} Q_k\{X_k(t)\} = \beta^t \widetilde{R}_j\{X_j(t)\} \prod_{k \neq j} Q_k\{X_k(t)\}. \tag{1}$$

The Q-multiplicative term in (1) ensures that no rewards are earned beyond any point at which Red has been killed. The quantity $\beta \in (0,1)$ is a discount factor and is included for generality. Provided we make natural model assumptions which guarantee that shooting stops (with Red dead or all Blues dead) after a finite number of engagements almost surely then we may also take $\beta = 1$ in what follows and consider undiscounted returns.

A policy is a rule for choosing actions at each $t = 0, 1, 2, \ldots$ in light of the history of the process to date. Under policy ν, use $\nu(t)$ for the choice made at t. We write the total expected return under policy ν as

$$E_\nu \left(\sum_{t=0}^{\infty} \beta^t \widetilde{R}_{\nu(t)}\{X_{\nu(t)}(t)\} \left[\prod_{j \neq \nu(t)} Q_j\{X_j(t)\} \right] \right). \tag{2}$$

The goal is to find policy ν^* to maximise the expected return in (2). The above is in the class of Markov decision models called *generalised bandits* introduced by Nash (1980). These models extend the multiarmed bandits of Gittins (1979,1989) by allowing a reward interdependence between the decision options, as expressed in the multiplicatively separable form to be found in (2). The theory of this class of processes has been developed in Glazebrook (1993), Glazebrook and Greatrix (1995) and Crosbie and Glazebrook (2000 a,b). For our purposes, the key fact is that for the class of problems outlined in (i) - (iv), there exists an optimal policy of *index form.* This is expressed in Theorem 1.

Theorem 1 (Nash(1980)) *There exist functions $G_j : \Omega_j \to \mathbb{R}^+$ such that, if Red is still alive at t then he optimally engages any Blue j^* for which*

$$G_{j^*}\{X_{j^*}(t)\} = \max_{1 \leq j \leq N} G_j\{X_j(t)\}. \tag{3}$$

The indices in (3) are broadly of Gittins type. To develop index $G_j(x)$ for some $x \in \Omega_j$, suppose that at $t = 0$, Blue j is in state x and is engaged continuously by Red. Let τ be a positive valued stopping time on the resulting process $\{X_j(t), t \geq 0\}$ which evolves from x according to the Markov law P_j. Use $\widetilde{R}_j(x, \tau)$ for the expected return earned during $[0, \tau)$ as expressed by

$$\widetilde{R}_j(x, \tau) = E\left[\sum_{t=0}^{\tau-1} \beta^t \widetilde{R}_j\{X_j(t)\} | X_j(0) = x \right], \tag{4}$$

and rewards are automatically terminated at Red's death. Develop a corresponding reward rate-like measure as

$$\widetilde{G}_j(x, \tau) = \widetilde{R}_j(x, \tau)(1 - E[\beta^\tau Q_j\{X_j(\tau)\} | X_j(0) = x])^{-1}. \tag{5}$$

The index $G_j(x)$ is the largest such reward rate, namely

$$G_j(x) = \sup_\tau \widetilde{G}_j(x, \tau). \tag{6}$$

A general methodology for index computation may be found in Glazebrook and Greatrix (1995).

We now present three particular models, each of which illustrate and present salient features of combat scenarios. In two cases, the indices which determine optimal engagement policies for Red are obtained in closed form. In the more complex "shoot-look-shoot" setup of Model 3, we give an algorithm for index development.

3 Model 1 - Red learns about the nature of Blue targets

While the general scenario (Red facing N Blues) is as above, we shall particularise to Model 1 in supposing that Blues come in B types and Red has imperfect information about the Blues he is facing. Note that "type" designation here may reflect any Blue characteristics which are relevant to determining outcomes as the conflict proceeds. Red's uncertainty about Blue is expressed through N independent prior distributions $\Pi^1, \Pi^2, \ldots, \Pi^N$ which summarise his beliefs before shooting starts. Hence Π_b^j is the probability that Red assigns to the event "Blue number j is of type b", $1 \leq j \leq N$, $1 \leq b \leq B$. At each time $t = 0, 1, 2, \ldots$ Red targets a single Blue and shooting continues until either Red is dead or all the Blues are. Conditional upon the event that a Blue targetted by Red is actually of type b, Red has a probability r_b of killing Blue while there is a probability θ_b that he himself is killed during the engagement. Red always has perfect information about whether each Blue is alive or dead and hence the model calls for the inclusion of state $\overline{\omega}_j$ within Ω_j as mentioned in Section 2(iii) above. All shooting outcomes are assumed independent. Should Red kill a type b Blue with his t^{th} shot then he receives a return $\beta^t R_b$. Red's goal is to maximise the expected return from Blues killed prior to his own destruction. The expectation concerned is taken both with respect to Red's prior beliefs as well as over realisations of the process. Note that the choices $\beta = 1$, $R_b = 1$, $1 \leq b \leq B$, lead to a maximisation of the number of Blues killed before Red's death.

A crucial feature of the model concerns Red's capacity to update his beliefs about the Blues he is facing in the light of past engagements by using Bayes' Theorem. In particular, if Blue j has been targetted in n engagements and he and Red have survived them all (note that these are the only event types of relevance for future decision-making) then the posterior distribution $\Pi^{j,n}$ summarising Red's updated beliefs about Blue j is given by

$$\Pi_b^{j,n} = \frac{\Pi_b^j (1-r_b)^n (1-\theta_b)^n}{\sum_{d=1}^B \Pi_d^j (1-r_d)^n (1-\theta_d)^n}, \quad 1 \leq b \leq B. \tag{7}$$

For notational simplicity, we shall refer to the denominator in (7) as $D_j(\Pi^j, n)$.

This problem may be represented within the general formulation of Section 2 (i)-(iv) as follows:

(i) State space Ω_j is taken to be $\mathbb{N} \cup \{\overline{\omega}_j\}$. If $X_j(t) = n \in \mathbb{N}$ then at time t, Blue j has been targetted in n engagements with Red, all of which have been inconclusive (neither killed).

(iii) Should action a_j be chosen at t when $X_j(t) = n$ then, following the resulting engagement a transition to $X_j(t+1)$ occurs according to Markovian law P_j where

$$P_j(n, n+1) = P(\text{neither Red nor Blue } j \text{ killed}) = D_j(\Pi^j, n+1)/D_j(\Pi^j, n);$$

$$P(n, \overline{\omega}_j) = P(\text{Blue } j \text{ killed but not Red}) = \sum_{b=1}^B \Pi_b^j r_b (1-r_b)^n (1-\theta_b)^{n+1} / D_j(\Pi^j, n),$$

and

$$P(n, \omega_j) = P(\text{Red killed}) = \sum_{b=1}^B \Pi_b^j \theta_b (1-r_b)^n (1-\theta_b)^n / D_j(\Pi^j, n).$$

The expected return (undiscounted) from the engagement in (iii) above is given by

$$R_j(n) = \sum_{b=1}^{B} \Pi_b^j R_b r_b (1-r_b)^n (1-\theta_b)^n / D_j(\Pi^j, n),\ n \in \mathbb{N}.$$

In following the prescription for index computation at the end of Section 2, note that in taking the supremum in (6), we may restrict to stationary stopping times i.e., those which stop the process $\{X_j(t), t \geq 0\}$ upon entry into a fixed *stopping set*. Hence, we consider the computation of index $G_j(n)$, appropriate for Blue j in state $n \in \mathbb{N}$. Specify positive integer r and write τ_r for Red's stopping time in which from time 0 (at which point $X_j(0) = n$), Red has r further engagements which target Blue j unless one or other of them is destroyed first. The random variable τ_r is the number of shots from Red which results from this, and cannot exceed r or be less than one. The expected return $\widetilde{R}_j(n, \tau_r)$ obtained by Red from these engagements is given by

$$\widetilde{R}_j(n, \tau_r) = \sum_{b=1}^{B} \Pi_b^j (1-r_b)^n (1-\theta_b)^n \left\{ \sum_{s=0}^{r-1} \beta^s R_b r_b (1-r_b)^s (1-\theta_b)^s \right\} / D_j(\Pi^j, n), \quad (8)$$

while we also have

$$E[\beta^{\tau_r} Q_j\{X_j(\tau_r)\} | X_j(0) = n] = \sum_{b=1}^{B} \Pi_b^j (1-r_b)^n (1-\theta_b)^n \times \left\{ \sum_{s=0}^{r-1} \beta^{s+1} r_b (1-r_b)^s (1-\theta_b)^{s+1} + \beta^r (1-r_b)^r (1-\theta_b)^r \right\} / D_j(\Pi^j, n). \quad (9)$$

From (5), (6), (8) and (9) and Theorem 1 we deduce the following:

Theorem 2 *If Red is still alive at t then he optimally targets any Blue j^* for which $X_{j^*}(t) \neq \overline{\omega}_{j^*}$ and such that*

$$G_{j^*}\{X_{j^*}(t)\} = \max_j G_j\{X_j(t)\},$$

where the maximisation is over those j for which $X_j(t) \neq \overline{\omega}_j$ and where

$$G_j(n) = \max_{r \geq 1} \left[\frac{\sum_{b=1}^{B} \Pi_b^j (1-r_b)^n (1-\theta_b)^n \left\{ \sum_{s=0}^{r-1} \beta^s R_b r_b (1-r_b)^s (1-\theta_b)^s \right\}}{\sum_{b=1}^{B} \Pi_b^j (1-r_b)^n (1-\theta_b)^n \{1 - F_{1b}(r) - F_{2b}(r)\}} \right], \quad n \in \mathbb{N},\ 1 \leq j \leq N, \quad (10)$$

where

$$F_{1b}(r) = \sum_{s=0}^{r-1} \beta^{s+1} r_b (1-r_b)^s (1-\theta_b)^{s+1},\ r \geq 1,\ 1 \leq b \leq B,$$

and

$$F_{2b}(r) = \beta^r (1-r_b)^r (1-\theta_b)^r,\ r \geq 1,\ 1 \leq b \leq B.$$

In order to understand index structure, introduce the "one-step index" $H_j(n)$ obtained by taking $r = 1$ in (10) as

$$H_j(n) = \frac{\sum_{b=1}^{B} \Pi_b^j (1-r_b)^n (1-\theta_b)^n R_b r_b}{\sum_{b=1}^{B} \Pi_b^j (1-r_b)^n (1-\theta_b)^n \{1-\beta+\beta\theta_b\}}. \tag{11}$$

It is straightforward to establish the following:

(a) If $H_j(n)$ is decreasing in n then the maximum in (10) is attained at $r = 1$ for all n and it then follows that $G_j(n) = H_j(n)$, $n \in \mathbb{N}$. If this behaviour holds good for all Blues then Red's optimal shooting policy is quasi-myopic (a one-step look ahead rule). Here indices decrease through to Blue's destruction and consequently the optimal index policy will tend to involve Red making frequent changes to the Blue targetted;

(b) If $H_j(n)$ is increasing in n, then the maximum in (10) is attained for all n in the limit as $r \to \infty$. When this happens the index $G_j(n)$ will take the form

$$\frac{\sum_{b=1}^{B} \Pi_b^j (1-r_b)^n (1-\theta_b)^n R_b r_b \{1-\beta(1-r_b)(1-\theta_b)\}^{-1}}{\sum_{b=1}^{B} \Pi_b^j (1-r_b)^n (1-\theta_b)^n \{(1-\beta+\beta\theta_b)[1-\beta(1-r_b)(1-\theta_b)]^{-1}\}}$$

and will be increasing in n. If this behaviour holds good for all Blues then Red, will in an optimal policy, persist in targeting individual Blues in turn until each is destroyed.

(c) If there are just two Blue types ($B = 2$), then it can be shown that one of the cases described in (a) and (b) must hold for each Blue target.

Lemma 3 *If $B = 2$ then either each $H_j(n)$ is increasing in n or each $H_j(n)$ is decreasing in n.*

Proof It is straightforward to show algebraically that, for any j and $n \in \mathbb{N}$

$$H_j(n) \geq H_j(n+1)$$

if and only if

$$\begin{aligned}[\{1-\beta+\beta\theta_1\}R_2 r_2 - \{1-\beta+\beta\theta_2\}R_1 r_1](1-r_1)(1-\theta_1) \\ \geq [\{1-\beta+\beta\theta_1\}R_2 r_2 - \{1-\beta+\beta\theta_2\}R_1 r_1](1-r_2)(1-\theta_2).\end{aligned}$$

This condition depends upon neither j nor n. The result follows. □

Comments

The one-step index $H_j(n)$ in (11) may be thought of (somewhat crudely) as a weighted average (with respect to the posterior distribution) of a *return/exposure index*

$$R_b r_b \{1-\beta+\beta\theta_b\}^{-1}$$

for Blues of type b. This index is high when R_b and r_b are large and when θ_b is small. It is plainly such Blue types which Red should target early. Note the dependence of this

quantity on θ_b. Plainly, Red should avoid targetting Blues with large associated θ-values as such engagements are high risk for him and his early demise will preempt the possibility of accumulating further returns.

The intuition behind Lemma 3 is that when $B = 2$, then for a specific Blue, as the number of inconclusive engagements increases, the balance of Red's beliefs about that Blue will move systematically *either* from it being of type 1 toward it being type 2 *or* in the opposite direction. One of these directions will yield an increasing index and one a decreasing index, depending on whether type 1 or type 2 has the larger return/exposure index.

4 Model 2 - Red inflicts accumulating damage upon Blue

The model discussed here is rather different in character to those of Sections 3 and 5. While there is now no Bayesian learning for Red, we do allow the N Blues targetted by Red to suffer accumulating damage during successive engagements. This is a step in the direction of shooting problems with targets whose characteristics evolve dynamically. See the comments in Section 7(b). We shall here make the simplifying assumption that an engagement consists of a shot by Red at Blue j, say, followed by a retaliatory strike from the Blue targetted. Further, a severely damaged Blue will be less lethal to Red. Should a Blue's damage be sufficient, it is deemed to have been killed. To express this, we assume that each Blue can be in any one of K states, labelled $\{1, 2, \ldots, K\}$ and that this state is observable without error by Red. As state k runs from 1 to K it represents increasing degrees of damage with $K = \overline{\omega}_j$ corresponding to Blue's death. The Markovian law P^j determines how Blue j evolves to higher damage states under successive attacks from Red, while $\theta_j(k)$ is the probability that Blue j can kill Red with a shot when in damage state k, where $P^j_{kl} = 0$, $l < k$, and $\theta_j(K) = 0$.

The general formulation of Section 2 (i)-(iv) should be adapted to this case as follows:

(i) State space Ω_j is $\{1, 2, \ldots, K\}$ with $K = \overline{\omega}_j$.

(iii) Should action a_j be chosen at t when $X_j(t) = k \in \{1, 2, \ldots, K-1\}$ then, following the resulting engagement between Red and Blue j a transition to $X_j(t+1)$ occurs according to Markovian law P_j where

$$P_j(k,l) = P(\text{engagement inconclusive, with Blue's damage } k \to l)$$

$$= P^j_{kl}\{1 - \theta_j(l)\},\ k \le l \le K-1;$$

$$P_j(k, \overline{\omega}_j) = P(\text{Blue killed but not Red}) = P^j_{kK};$$

and

$$P_j(k, \omega_j) = P(\text{Red killed}) = \sum_{l=k}^{K-1} P^j_{kl}\theta_j(l).$$

(iv) The expected return from the engagement in (iii) above is given by

$$R_j(k) = \beta R_j P^j_{kK},\ k \in \{1, 2, \ldots, K-1\},$$

where we assume that the reward R_j is received when Blue j enters state K.

We consider the computation of index $G_j(k)$, appropriate for calibrating Blue j when in state $k \in \{1, 2, \ldots, K-1\}$. To this end, suppose that $X_j(0) = k$ and that Blue j is subjected to successive engagements with Red. Any stationary positive-valued stopping time τ on Blue's evolving state corresponds to a choice of subset $S(k) \subseteq \{k, k+1, \ldots, K-1\}$ such that

$$\tau = \min[t; t > 0 \text{ and } X_j(t) \in S(k) \cup \{\overline{\omega}_j\} \cup \{\omega_j\}]. \tag{12}$$

The expected return $\widetilde{R}_j(k, \tau)$ obtained by Red in all engagements with Blue j up to stopping time τ is given by

$$\widetilde{R}_j(k, \tau) = R_j Z_k^j\{S(k)\};$$

where the quantities $\{Z_l^j\{S(k)\},\ 1 \leq l \leq K-1\}$ satisfy recursions

$$Z_l^j\{S(k)\} = \beta P_{lK}^j + \beta \sum_{m \notin S(k)} P_{lm}^j\{1 - \theta_j(m)\} Z_m^j\{S(k)\},\ 1 \leq l \leq K-1.$$

The corresponding reward rate from (5) is given by

$$\widetilde{G}_j(k, \tau) = R_j Z_k^j\{S(k)\}[1 - Z_k^j\{S(k)\}]^{-1}.$$

In Theorem 4, we use $2^{\{k,k+1,\ldots,K-1\}}$ for the collection of subsets of $\{k, k+1, \ldots, K-1\}$.

Theorem 4 *If Red is still alive at t then he optimally targets any Blue j^* for which $X_{j^*}(t) \neq \overline{\omega}_{j^*}$ and such that*

$$G_{j^*}\{X_{j^*}(t)\} = \max_j G_j\{X_j(t)\}$$

where the maximisation is over those j for which $X_j(t) \neq \overline{\omega}_j$ and where

$$G_j(k) = \max_{S(k)}(R_j Z_k^j\{S(k)\}[1 - Z_k^j\{S(k)\}]^{-1}),\ k \in \{1, 2, \ldots, K-1\},\ 1 \leq j \leq N, \tag{13}$$

where the maximisation in (13) is over $2^{\{k,k+1,\ldots,K-1\}}$.

While there are efficient algorithms for computing the indices in (13) (including the adaptive greedy algorithm of Robinson (1982) or the "restart-in-k" construction of Katehakis and Veinott (1987)) we now introduce plausible assumptions regarding the system's stochastic structure which greatly simplify index structure.

Assumptions

(1) For all j, $\sum_{l=m}^K P_{kl}^j$ is increasing in k for each choice of $m \in \{k, k+1, \ldots, K\}$;

(2) For all j, $\theta_j(k)$ is decreasing in k.

Assumption (1) states that in the engagement discussed in (iii) above, Blue's new damage state $X_j(t+1)$ is stochastically increasing in its old damage state, $X_j(t)$. Assumption (2) states that Blue j becomes less lethal to Red as it is increasingly damaged.

We proceed to develop index $G_j(k)$ by introducing quantities

$$Z_l^j = Z_l^j(\phi) = \beta P_{lK}^j + \beta \sum_{m=l}^{K-1} P_{lm}^j \{1 - \theta_j(m)\} Z_m^j$$

$$= \beta \sum_{m=l}^{K} P_{lm}^j \{1 - \theta_j(m\} Z_m^j, \text{ where } Z_K^j = 1. \tag{14}$$

Lemma 5 *The quantity Z_k^j is increasing in k, for each j, $1 \leq j \leq N$.*

Proof It is plain that $Z_{K-1}^j \leq Z_K^j = 1$. We shall proceed by induction, will suppose that $Z_{k+1}^j \leq Z_{k+2}^j \leq \ldots \leq Z_K^j$ and will show that the inequality $Z_k^j \leq Z_{k+1}^j$ follows.

First observe that, from Assumption (2) and the inductive hypothesis, it follows that

$$\{1 - \theta_j(k+1)\} Z_{k+1}^j \leq \{1 - \theta_j(k+2)\} Z_{k+2}^j \leq \ldots \leq \{1 - \theta_j(K)\} Z_K^j = 1.$$

Now, utilising Assumption (1) we have that

$$\begin{aligned} Z_{k+1}^j &= \beta \sum_{l=k+1}^{K} P_{k+1l}^j \{1 - \theta_j(l)\} Z_l^j \\ &= \beta \{1 - \theta_j(k+1)\} Z_{k+1}^j \\ &\qquad + \beta \sum_{l=k+2}^{K} [\{1 - \theta_j(l)\} Z_l^j - \{1 - \theta_j(l-1)\} Z_{l-1}^j] (\sum_{m=l}^{K} P_{k+1,m}^j) \\ &\geq \beta \{1 - \theta_j(k+1)\} Z_{k+1}^j \\ &\qquad + \beta \sum_{l=k+2}^{K} [\{1 - \theta_j(l)\} Z_l^j - \{1 - \theta_j(l-1)\} Z_{l-1}^j] (\sum_{m=l}^{K} P_{k,m}^j). \end{aligned} \tag{15}$$

Similarly, we have that

$$\begin{aligned} Z_k^j &= \beta \sum_{l=k}^{K} P_{kl}^j \{1 - \theta_j(l)\} Z_l^j \\ &= \beta P_{kk}^j \{1 - \theta_j(k)\} Z_k^j + \beta \{1 - \theta_j(k+1)\} Z_{k+1}^j (\sum_{l=k+1}^{K} P_{k,l}^j) \\ &\qquad + \beta \sum_{l=k+2}^{K} [\{1 - \theta_j(l)\} Z_l^j - \{1 - \theta_j(l-1)\} Z_{l-1}^j] (\sum_{m=l}^{K} P_{k,m}^j). \end{aligned} \tag{16}$$

From (15) and (16) we infer that

$$Z_{k+1}^j [1 - \beta \{1 - \theta_j(k+1)\}] \geq Z_k^j [1 - \beta P_{kk}^j \{1 - \theta_j(k)\}] + Z_{k+1}^j \beta \{1 - \theta_j(k+1)\} (1 - P_{kk}^j)$$

and hence that

$$Z_{k+1}^j [1 - \beta P_{kk}^j \{1 - \theta_j(k+1)\}] \geq Z_k^j [1 - \beta P_{kk}^j \{1 - \theta_j(k)\}]. \tag{17}$$

We now use $\theta_j(k) \geq \theta_j(k+1)$ together with (17) to infer that $Z_k^j \leq Z_{k+1}^j$. The induction goes through and the proof is concluded. □

Theorem 6 *Under Assumptions (1, 2), Blue index $G_j(k)$ is given by*

$$G_j(k) = R_j Z_k^j (1 - Z_k^j)^{-1},\ k \in \{1, 2, \ldots, K-1\},\ 1 \leq j \leq N,$$

and is increasing in k.

Proof Suppose that $X_j(0) = k \in \{1, 2, \ldots, K-1\}$ and that stopping time τ has associated stopping set $S(k)$ as in (12), for which $P\{X_j(\tau) \in S(k)\} > 0$. Use $\widetilde{\tau}$ for the stopping time corresponding to $S(k) = \phi$. Plainly $\tau \leq \widetilde{\tau}$ with probability one. Utilising the above defined quantities we have that

$$\begin{aligned} G_j(k, \widetilde{\tau}) &= R_j Z_k^j (1 - Z_k^j)^{-1} \\ &= \frac{R_j Z_k^j \{S(k)\} + E[\beta^\tau I\{X_j(\tau) \in S(k)\} R_j Z_{X_j(\tau)}^j]}{1 - Z_k^j \{S(k)\} + E(\beta^\tau I\{X_j(\tau) \in S(k)\}[1 - Z_{X_j(\tau)}^j])}. \end{aligned} \tag{18}$$

But from Lemma 5 we have that,

$$X_j(\tau) \in S(k) \Rightarrow R_j Z_{X_j(\tau)}^j \{1 - Z_{X_j(\tau)}^j\}^{-1} \geq R_j Z_k^j \{1 - Z_k^j\}^{-1} = G_j(k, \widetilde{\tau}). \tag{19}$$

From (18) and (19) it follows that

$$G_j(k, \widetilde{\tau}) \geq \frac{R_j Z_k^j \{S(k)\} + G_j(k, \widetilde{\tau}) E(\beta^\tau I\{X_j(\tau) \in S(k)\}[1 - Z_{X_j(\tau)}^j])}{1 - Z_k^j \{S(k)\} + E(\beta^\tau I\{X_j(\tau) \in S(k)\}[1 - Z_{X_j(\tau)}^j])}. \tag{20}$$

It now follows immediately from (20) that

$$G_j(k, \widetilde{\tau}) \geq R_j Z_k^j \{S(k)\}[1 - Z_k^j \{S(k)\}]^{-1} = G_j(k, \tau) \tag{21}$$

for any τ and associated stopping set $S(k)$. The result immediately follows from (21) and the form of the index $G_j(k)$ given in (13). The increasing nature of $G_j(k)$ follows from the increasing nature of Z_k^j, reported in Lemma 5. □

Comments

(a) Under Assumptions (1,2), the increasing nature of index $G_j(k)$ in k means that in an optimal policy Red will engage each Blue continually until the latter is killed (unless Red dies first). This approach is intuitive since Blue's accumulating damage through his engagements not only brings his own death closer (Assumption (1)), but also makes him progressively less lethal to Red (Assumption (2)). Hence it is clear that Red should continue shooting at a partly damaged Blue and the index policy guarantees that this is so.

(b) To see how the index $G_j(k)$ depends upon Blue j's lethality, consider two extreme cases. Suppose first that Blue j is lethal right up to its own destruction, namely

$$\theta_j(l) \cong 1, 1 \leq l \leq K-1.$$

It then follows that

$$Z_k^j \cong \beta P_{k,K}^j$$

and hence that

$$G_j(k) \cong \beta R_j P^j_{k,K}\{1 - \beta P^j_{k,K}\}^{-1}. \tag{22}$$

Any shot by Red at such a Blue is a gamble that the latter will be killed with a single shot. Suppose now that Blue j poses very little retaliatory threat to Red in that

$$\theta_j(l) \cong 0,\ 1 \leq l \leq K-1.$$

Consider the quantities $\{\widetilde{Z}^j_l, 1 \leq l \leq K-1\}$ satisfying the recursions

$$\widetilde{Z}^j_l = \beta P^j_{lK} + \beta \sum_{m=l}^{K} P^j_{lm} \widetilde{Z}^j_m, 1 \leq l \leq K-1,\ \widetilde{Z}^j_K = 1.$$

We now have

$$G_j(k) \cong R_j \widetilde{Z}^j_k (1 - \widetilde{Z}^j_k)^{-1} \tag{23}$$

and Red's only concern now is the speed with which Blue j can be killed and the return R_j claimed. Not surprisingly, the index in (22) will be smaller than that in (23).

5 Model 3 - 'Shoot-look-shoot' for Red

Our goal here is to give the reader insight concerning the generality of our modelling/solution approach by introducing developments of Model 1 of considerable practical import. The general scenario and Π^j_b, R_b, r_b and β are all as before. However, now we shall suppose that after every shot by Red, the targetted Blue is inspected and categorised (with error) according to Blue target type and alive/dead. Write $\delta \in \{1, 2, \ldots, B\} \times \{\text{alive,dead}\}$ for a generic classification. We have that

$$P[\text{Blue judged to be } \delta | \text{Blue is alive of type } b] = \phi_{\delta b}$$
$$P[\text{Blue judged to be } \delta | \text{Blue is dead of type } b] = \phi_{\delta \bar{b}}$$

where $1 \leq b \leq B$. We shall also suppose that Red's vulnerability depends upon whether the targetted Blue is alive or dead. We use θ_b for the probability that Red is killed during an engagement in which he targets a Blue of type b who is still alive. This becomes $\overline{\theta}_b$ if the targetted Blue is dead.

Red now gathers information about the Blues he is facing through the series of engagements in a more complicated way than for Model 1. Index policies will remain optimal, but the index structure will be more complex and simple closed forms as in (10) and Theorem 6 above should not be expected. Consider Blue target j with assigned prior Π^j. At time t, if Red is still alive then sufficient statistics from the history of Red's past engagements which targetted Blue j which determine Red's posterior distribution for this Blue are:

(a) the number of engagements targetting Blue j (n);

(b) the outcomes of Red's subsequent inspections ($\boldsymbol{\delta} = \{\delta_1, \delta_2, \ldots, \delta_n\}$).

We take these sufficient statistics as Blue j's state at t while Red is alive and write $X_j(t) = (n, \boldsymbol{\delta})$. Red's posterior probability, given this history, that Blue j is of type b and is still alive is proportional to

$$\Pi_b^j (1-r_b)^n (1-\theta_b)^n \left(\prod_{l=1}^{n} \phi_{\delta_l b} \right) \equiv \Pi_b^j P_b(n, \boldsymbol{\delta}) \equiv \Pi_b^j P_b\{X_j(t)\}. \tag{24}$$

Red's posterior probability, given this history, that Blue j is of type b but is now dead is proportional to

$$\Pi_b^j \sum_{k=1}^{n} (1-r_b)^{k-1} r_b (1-\theta_b)^k (1-\overline{\theta}_b)^{n-k} \left(\prod_{l=1}^{k-1} \phi_{\delta_l b} \right) \left(\prod_{l=k}^{n} \phi_{\delta_l \overline{b}} \right)$$
$$\equiv \Pi_b^j \overline{P}_b(n, \boldsymbol{\delta}) \equiv \Pi_b^j \overline{P}_b\{X_j(t)\}, \tag{25}$$

as before. Hence, given the history summarised by $X_j(t)$, Red's posterior probabilities for Blue j are given by

$$P[\text{Blue } j \text{ is alive and of type } b | X_j(t)] = \frac{\Pi_b^j P_b\{X_j(t)\}}{\sum_{d=1}^{B} \Pi_d^j [P_d\{X_j(t)\} + \overline{P}_d\{X_j(t)\}]}, \quad 1 \le b \le B, \tag{26}$$

and

$$P[\text{Blue } j \text{ is dead and of type } b | X_j(t)] = \frac{\Pi_b^j \overline{P}_b\{X_j(t)\}}{\sum_{d=1}^{B} \Pi_d^j [P_d\{X_j(t)\} + \overline{P}_d\{X_j(t)\}]}, \quad 1 \le b \le B. \tag{27}$$

Our scheduling problem may be represented within the formulation of Section 2 (i)-(iv) as follows:

(i) State space Ω_j is the set of all possible histories $(n, \boldsymbol{\delta})$. Since in general under this model, Red can never be certain that Blue j has been killed, there is no state $\overline{\omega}_j$.

(ii) Should action a_j be chosen at t when $X_j(t) = (n, \boldsymbol{\delta})$ there are two modes of transition to $X_j(t+1)$, depending upon whether Red is killed in the engagement or not. If Red is not killed, we have a state transition of the form

$$(n, \boldsymbol{\delta}) = X_j(t) \to X_j(t)(\delta) \equiv \{n+1, (\boldsymbol{\delta}, \delta)\} \tag{28}$$

with probability

$$\frac{\sum_{b=1}^{B} \Pi_b^j [P_b\{X_j(t)\}\{(1-r_b)(1-\theta_b)\phi_{\delta b} + r_b(1-\theta_b)\phi_{\delta \overline{b}}\} + \overline{P}_b\{X_j(t)\}(1-\overline{\theta}_b)\phi_{\delta \overline{b}}]}{\sum_{b=1}^{B} \Pi_b^j [P_b\{X_j(t)\} + \overline{P}_b\{X_j(t)\}]}. \tag{29}$$

If Red is killed then $X_j(t+1) = \omega_j$. This happens with probability

$$\frac{\sum_{b=1}^{B} \Pi_b^j [P_b\{X_j(t)\}\theta_b + \overline{P}_b\{X_j(t)\}\overline{\theta}_b]}{\sum_{b=1}^{B} \Pi_b^j [P_b\{X_j(t)\} + \overline{P}_b\{X_j(t)\}]}. \tag{30}$$

In order to develop the indices which determine optimal shooting policies for Red, it will assist notationally if we drop the Blue target identifier j and use H for a generic sufficient history of the form $(n, \boldsymbol{\delta})$ above. We wish to obtain $G(\Pi, H)$, namely the index for a Blue with prior Π and history H. We shall use an adapted version of the "restart-in-H" approach to index computation proposed by Katehakis and Veinott (1987). See also Glazebrook and Greatrix (1995). To this end, use $\Omega(H)$ for the set of histories reachable (in the obvious sense) from history H and $B\{\Omega(H)\}$ for the set of bounded real-valued functions on $\Omega(H)$.

The "restart-in-H" problem appropriate for index computation is a Markov Decision Problem with initial state H. Actions at each stage of the process are *either* that Red should engage with Blue *or* that the current state should be reset to H and that Red should then engage with Blue. Transition probabilities and returns are as above. The process is terminated by Red's death at any stage. From Katehakis and Veinott (1987) we infer that $G(\Pi, H)$ is the value function for this problem. To obtain $G(\Pi, H)$ we use the following iterative scheme: let $u \in B\{\Omega(H)\}$ and $H' \in \Omega(H)$.

Consider the transform $T_H : B\{\Omega(H)\} \to B\{\Omega(H)\}$ given by

$$\{T_H(u)\}(H') = \max\left\{\left(\sum_{b=1}^{B}\Pi_b P_b(H')\left[R_b r_b + \beta r_b(1-\theta_b)\sum_{\delta}\phi_{\delta\bar{b}}u\{H'(\delta)\}\right.\right.\right.$$
$$\left.\left.\left.+\beta(1-r_b)(1-\theta_b)\sum_{\delta}\phi_{\delta b}u\{H'(\delta)\}\right] + \sum_{b=1}^{B}\Pi_b\overline{P}_b(H')\beta(1-\overline{\theta}_b)\sum_{\delta}\phi_{\delta\bar{b}}u\{H'(\delta)\}\right)\right.$$
$$\left.\times\left(\sum_{b=1}^{B}\Pi_b\{P_b(H')+\overline{P}_b(H')\}\right)^{-1};\right.$$
$$\left(\sum_{b=1}^{B}\Pi_b P_b(H)\left[R_b r_b + \beta r_b(1-\theta_b)\sum_{\delta}\phi_{\delta\bar{b}}u\{H(\delta)\} + \beta(1-r_b)(1-\theta_b)\sum_{\delta}\phi_{\delta b}u\{H(\delta)\}\right]\right.$$
$$\left.\left.+\sum_{b=1}^{B}\Pi_b\overline{P}_b(H)\beta(1-\overline{\theta}_b)\sum_{\delta}\phi_{\delta\bar{b}}u\{H(\delta)\}\right)\left(\sum_{b=1}^{B}\Pi_b\{P_b(H)+\overline{P}_b(H)\}\right)^{-1}\right\}. \quad (31)$$

We now use T_H^n for an n-fold application of T_H, namely

$$T_H^1 = T_H \text{ and } T_H^n = T_H(T_H^{n-1}),\ n \geq 2.$$

Standard results concerning value iteration for discounted Markov Decision Processes (see, for example, Ross (1970)) yield the value function for the restart-in-H problem as

$$\lim_{n\to\infty}\{T_H^n(u)\}(H) = G(\Pi, H) \text{ for all } u \in B\{\Omega(H)\}. \quad (32)$$

Theorem 7 summarises the results of the above analysis for this case.

Theorem 7 *If Red is still alive at t then he optimally targets any Blue j^* for which*

$$G_{j^*}\{\Pi^{j^*}, X_{j^*}(t)\} = \max_{1\leq j\leq N} G_j\{\Pi^j, X_j(t)\},$$

where the indices are determined by the iterative scheme in (31) and (32).

6 Numerical Study

We report on the outcome of a simulation study whose aim is to explore statistical properties of the optimal (index) shooting policy and other competitor policies for Red. This study will be in the context of instances of a minor variant of Model 1 for which $N = 10$ (ten Blue targets) and $B = 5$ (five Blue types). In this variant we suppose that Red cannot be killed in any engagement in which he kills a Blue opponent. Table 1 contains details of the Blue types. The reader will observe that model parameters have been chosen such that the Blues which yield highest rewards for Reds are the most difficult to kill. Targetting these also makes Red more vulnerable. Red must strike an optimal balance between garnering returns from Blue kills and remaining alive.

b	r_b	θ_b	R_b
1	0.9	0.2	50
2	0.7	0.3	125
3	0.5	0.4	250
4	0.3	0.5	500
5	0.1	0.6	1000

Table 1: Details of the five Blue types

The study consisted of 40,000 runs - with 10,000 runs being conducted under each of four different policies for Red. For each run, the 50 probabilities Π_b^j are drawn independently from a $U(0,1)$ distribution and normalised to achieve

$$\sum_{b=1}^{5} \Pi_b^j = 1, \ 1 \leq j \leq 10.$$

Discount rate β was set equal to 0.95 in all cases. The four shooting policies for Red are as follows:

(I) **Index Policy** - This is the policy which maximises the expected return earned by Red before his own death;

(II) **Myopic Policy** - Here Red's policy is to shoot next at whichever Blue is still alive and offers him the highest one-stage return. If Blue j has prior Π^j and has had n inconclusive engagements with Red to date, this one-stage return is given by

$$\left\{\sum_{b=1}^{B} \Pi_b^j (1-r_b)^n (1-\theta_b)^n R_b r_b\right\} \left\{\sum_{b=1}^{B} \Pi_b^j (1-r_b)^n (1-\theta_b)^n\right\}^{-1};$$

(III) **Random Policy** - At each stage, Red chooses between the still-alive Blues at random, with all Blue targets equally likely;

(IV) **Round Robin Policy** - Red cycles around the Blue targets (which are still alive) in numerical order. The first target is chosen at random.

For each policy, Tables 2 and 3 contain summaries of the 10,000 runs conducted. Table 2 gives a statistical summary of the returns earned by Red and records the mean return, the minimum (Min), lower quartile (LQ), median (Med), upper quartile (UQ) and maximum (Max). Table 3 gives a similar summary for the number of Blue targets destroyed by Red before he himself is killed. The final column of Table 3 also gives, for each policy, the percentage of runs for which Red is killed (i.e. before all the Blues are).

Policy	Mean	Min	LQ	Med	UQ	Max
Index	402.53	0.00	50.00	250.00	600.00	3208.32
Myopic	351.25	0.00	0.00	125.00	500.00	3393.94
Random	359.36	0.00	0.00	168.75	525.00	3353.82
Round-Robin	363.73	0.00	0.00	168.75	546.52	3476.21

Table 2: Summary of Red's returns using four different shooting policies

Policy	Mean	Min	LQ	Med	UQ	Max	% Red killed
Index	2.41	0	1	2	4	10	99.23
Myopic	1.52	0	0	1	2	10	98.90
Random	1.91	0	0	1	3	10	99.01
Round-Robin	1.87	0	0	1	3	10	99.24

Table 3: Summary of number of Blues killed and Red's death rate for four shooting policies for Red.

That the index policy should outperform the others with regard to its mean total return is guaranteed by Theorem 1. What is of note from the numerical results is its comprehensive dominance of the alternatives studied with regard to all summary measures of returns obtained and targets killed. The poor performance of the myopic policy is rooted in its indifference to the issue of Red's vulnerability when targetting different Blues. Its very low median return (half that of the index policy) speaks of many conflicts in which Red is killed very early. The evidence from the study is that in this context it would be better for Red to shoot at random (or in a round robin fashion) than myopically. In addition to maximising returns, the index policy also outperforms the others with regard to numbers of Blues killed - see Table 3. The probability that Red does not survive the conflict is roughly policy independent.

7 Extensions and comments

A range of extensions to the models discussed in Sections 2-5 is possible for which index policies either remain optimal or (at least) continue to perform well. See Gaver, Glazebrook and Pilnick (1991) for a discussion of such model elaborations in a different problem context and Glazebrook, Gaver and Jacobs (2001) for a discussion which focusses specifically on a variant of Model 1. Important among such model developments are those which acknowledge that Red has a finite number of bullets only. We note the following:

(a) If Red has a finite number of bullets then we have "finite horizon" versions of the (potentially infinite) scenarios analysed in preceding sections. The index policies developed there will continue to be optimal for Red in the so-called deteriorating cases in which each Blue's index decreases almost surely after each inconclusive engagement. This will happen, for example, in Model 1 when $H_j(n)$ (see (11)) is decreasing in n for each j. Note from Lemma 3 that when $B = 2$, the H_j are guaranteed to be either all increasing or all decreasing. The proof contains a condition under which the decreasing case is guaranteed. Other versions of the "finite horizon" problem outside of the deteriorating case are not indexable in general, but index policies will usually continue to perform very well. Mitchell (2003) has conducted a numerical study of the performance of index policies for a version of Model 2 in which the Blues do not retaliate. In the interests of brevity, we shall omit further details other than to point out that the broad approach to the numerical investigation was as in the study outlined in Section 6. For scenarios in which Red was limited in the numerical investigation to 25, 50 and 100 bullets respectively and for which Red had an infinite supply, his expected return from killing Blues was estimated for four shooting policies. These policies broadly correspond to those considered in Section 6. In Table 4 below, find values of $100\{(R^{INDEX} - R^{POLICY})/R^{INDEX}\}$, where we use R^{INDEX}, R^{POLICY} for the total expected returns for Red under the index policy and under any specified shooting policy respectively. Please note the outstandingly strong performance of the index policy in the short horizon (25 bullet) case. By the time Red is assumed to have 50 bullets, the profile of returns is much as in the infinite horizon case. The reader should note that, in contrast to the results in Tables 2 and 3, different lethalities of the Blues are no longer present to undermine the performance of the myopic policy.

	Number of bullets available to Red			
Policy	25	50	100	∞
Myopic	7.84%	5.78%	5.73%	5.80%
Random	60.43%	44.56%	43.83%	43.95%
Round-robin	38.00%	29.51%	29.07%	29.02%

Table 4: Percentage return lost when Red implements a shooting policy other than the index policy.

(b) In Models 1 and 3, each Blue target is supposed to have fixed characteristics, which however may be imperfectly known by Red. In Model 1 these are summarised by the triple (r_b, θ_b, R_b) for Blues of type b. In Model 2 target characteristics are dynamic, and change by means of a process of accumulating damage caused by Red's shots. The approach described in the paper and the general model of Section 2 can easily accommodate evolution of Blue target characteristics during targetting by Red. However, we may wish to model target dynamics while not underfire. For example, Blue may wish to re-deploy alive targets not under current fire so as to be more lethal to Red. This possibility takes us into a class of decision processes which are a generalised form of the restless bandit problems of Whittle (1988). While restless bandit problems are intractable in general, Whittle proposed an index heuristic (well defined under given conditions). These index heuristics have proved outstandingly effective in other

application contexts. See, for example, Glazebrook, Lumley and Ansell (2003) and Glazebrook and Mitchell (2002). The first author is conducting an extensive research programme in this challenging yet important area.

In closing, we briefly consider issues for the Blue force. For definiteness, the discussion will again be conducted in the context of Model 1, discussed in Section 3. A natural first question for the controller of Blue concerns what force he needs to deploy to destroy an optimally shooting Red with a given large probability $1-\varepsilon$. This turns out to be straightforward to assess. Suppose that N_b Blues of type b are deployed, $1 \leq b \leq B$. The probability of Red's ultimate survival (having destroyed all Blues) does not depend upon his strategy for engaging them. Hence, we may suppose that Red targets each Blue in a continuous set of exchanges until one or both are destroyed. In such an engagement it is easy to show that

$$P(\text{Red survives and kills Blue of type } b) = \frac{r_b(1-\theta_b)}{r_b+\theta_b(1-r_b)} \equiv \Psi_b,\ 1 \leq b \leq B,$$

$$P(\text{Red is killed and Blue of type } b \text{ survives}) = \frac{(1-r_b)\theta_b}{r_b+\theta_b(1-r_b)} \equiv \Delta_b,\ 1 \leq b \leq B,$$

and

$$P(\text{Red and Blue of type } b \text{ are both killed}) = \frac{r_b\theta_b}{r_b+\theta_b(1-r_b)} \equiv \Theta_b,\ 1 \leq b \leq B.$$

Hence the probability that Red survives the battle with N_b type b Blues, $1 \leq b \leq B$, is given by

$$\prod_{b=1}^{B} \Psi_b^{N_b}$$

and this is required to be no greater than ε.

If we now ask how the Blue force should accomplish the destruction of Red with given probability at least cost to itself, then Red's shooting strategy does come into play since, for example, Red may tend to target "expensive" Blues first. Now, suppose that Red shoots optimally with a single weapon and consider a simple scenario for Model 1 in which $B=2$ and all indices are increasing in n. See Lemma 3. Hence Red's optimal policy targets each Blue continuously until one or other is destroyed. Write $C(N_1,N_2)$ for the expected cost to the Blue force of the deployment of N_b type b's, $b=1,2$, against an optimally shooting Red. We shall assume here that $\beta=1$. Blue's optimisation problem is

$$\begin{aligned} \min_{N_1,N_2} \quad & C(N_1,N_2) \\ \text{subject to} \quad & \Psi_1^{N_1}\Psi_2^{N_2} \leq \varepsilon. \end{aligned} \tag{33}$$

We now describe a scenario in which $C(N_1,N_2)$ may be computed easily. Suppose that Red's prior distributions for the Blues he faces are obtained by moderating initial ignorance about them (expressed by $P(\text{Blue is supposed to be of type } b)= 0.5,\ b=1,2$) by means of information obtained from a sensor. This sensor can only judge Blue type with error. We have

$$P(\text{Blue judged to be of type } b_1 | \text{Blue is of type } b_2) = \phi_{b_1 b_2}$$

for all choices of b_1, b_2. Hence Red allocates to each Blue one of two possible priors Π^b, $b = 1, 2$ according to the judgement of the sensor. We have

$$\Pi_1^b = P(\text{Blue is of type 1}|\text{Blue judged to be of type } b) = \frac{\phi_{b1}}{\phi_{b1} + \phi_{b2}}.$$

Let X_1 be a binomial $Bin(N_1, \phi_{11})$ random variable representing the number of the N_1 type 1 Blues judged by the sensor to be of type 1 and hence given prior Π^1 by Red. Similarly $X_2 \sim Bin(N_2, \phi_{22})$. Red faces $X_1 + N_2 - X_2$ Blues to which he allocates prior Π^1 and initial index $G_1(0)$ and $N_1 - X_1 + X_2$ Blues to which he allocates prior Π^2 and initial index $G_2(0)$. Suppose that $G_1(0) > G_2(0)$ and so Red engages first all Blues judged to be of type 1. If Red faces two or more Blues with the same index, he chooses between them at random. Now write $c(b_1, b_2)$ for the expected cost to Blue when Red engages b_1 type 1 Blues and b_2 type 2 Blues in random order. If C_b is the cost of deploying a single Blue of type b, then

$$(b_1 + b_2)c(b_1, b_2) = b_1\{C_1(\Psi_1 + \Theta_1) + \Psi_1 c(b_1 - 1, b_2)\} + b_2\{C_2(\Psi_2 + \Theta_2) + \Psi_2 c((b_1, b_2 - 1)\},$$
$$c(0, 0) = 0$$

which enables recursive calculation of any $c(b_1, b_2)$. We deduce that the expected cost to Blue of the chosen deployment is given by

$$C(N_1, N_2) = E\left\{c(X_1, N_2 - X_2) + \Psi_1^{X_1}\Psi_2^{N_2 - X_2} c(N_1 - X_1, X_2)\right\}$$

and this may now be used in (33). In more complicated situations, Blue's expected cost may be computed via suitable development of the methodologies described by Bertsimas and Niño-Mora (1996) for multi-armed bandits.

Acknowledgements

The first two authors acknowledge the support provided by the Engineering and Physical Sciences Research Council, including that through grant GR/S45188/01

References

Barkdoll, T. C., Gaver, D. P., Glazebrook, K. D., Jacobs, P. A. & Posadas, S. (2002), 'Suppression of Enemy Air Defences (SEAD) as an Information Duel', *Nav. Res. Logist.* **49**, 723–742.

Bertsimas, D. & Niño-Mora, J. (1996), 'Conservation laws, extended polymatroids and multi-armed bandit problems: a polyhedral approach to indexable systems', *Math. Oper. Res.* **21**, 257–306.

Crosbie, J. H. & Glazebrook, K. D. (2000*a*), 'Evaluating policies for generalized bandits via a notion of duality', *J. Appl. Probab.* **37**, 540–546.

Crosbie, J. H. & Glazebrook, K. D. (2000*b*), 'Index policies and a novel performance space structure for a class of generalised branching bandit problems', *Math. Oper. Res.* **25**, 281–297.

Fay, N. A. & Walrand, J. C. (1991), 'On approximately index strategies for generalized arm problems', *J. Appl. Probab.* **28**, 602–612.

Gaver, D. P., Glazebrook, K. D. & Pilnick, S. E. (1991), 'Optimal sequential replenishment of ships during combat', *Nav. Res. Logist.* **38**, 637–668.

Gittins, J. C. (1989), *Multi-armed Bandit Allocation Indices*, Wiley, Chichester.

Gittins, J. C. (2004), 'Bandit processes and dynamic allocation indices (with discussion)', *J. Roy. Statist. Soc.* **B41**, 148–177.

Gittins, J. C. & Jones, D. M. (1974), A dynamic allocation index for the sequential design of experiments, *in* 'Progress in Statistics', J. Gani & I. Vince, eds, North-Holland, Amsterdam, pp. 241–266.

Glazbrook, K. D. & Mitchell, H. M. (2002), 'An index policy for a stochastic scheduling model with improving/deteriorating jobs', *Nav. Res. Logist.* **49**, 706–721.

Glazebrook, K. D. (1993), 'Indices for families of competing Markov decision processes with influence', *Ann. Appl. Probab.* **3**, 1013–1032.

Glazebrook, K. D., Gaver, D. P. & Jacobs, P. A. (2001), Military stochastic scheduling treated as a multi-armed bandit problem, Technical Report NPS-OR-01-010, Naval Postgraduate School, Monterey, CA.

Glazebrook, K. D. & Greatrix, S. (1995), 'On transforming an index for generalised bandit problems', *J. Appl. Probab.* **32**, 168–182.

Glazebrook, K. D., Lumley, R. R. & Ansell, P. S. (2003), 'Index heuristics for multi-class M/G/1 systems with nonpreemptive service and convex holding costs', *Queueing Syst.* **45**, 81–111.

Glazebrook, K. D. & Washburn, A. (2004), 'Shoot-look-shoot: A review and extension', *Oper. Res., (to appear)* .

Katehakis, M. N. & Veinott, A. F. (1987), 'The multi-armed bandit problem - decomposition and computation', *Math. Oper. Res.* **12**, 262–268.

Manor, G. & Kress, M. (1997), 'Optimality of the greedy shooting strategy in the presence of incomplete damage information', *Nav. Res. Logist.* **44**, 613–622.

Mitchell, H. M. (2003), PhD thesis, Newcastle University, Newcastle upon Tyne, UK.

Nash, P. (1980), 'A generalised bandit problem', *J. Roy. Statist. Soc.* **B42**, 165–169.

Robinson, D. (1982), 'Algorithms for evaluating the dynamic allocation index', *Oper. Res. Lett.* **1**, 72–74.

Ross, S. M. (1970), *Applied probability models with optimization applications*, Holden-Day, San Francisco.

Weber, R. R. (1992), 'On the Gittins index for multi-armed bandits', *Ann. Appl. Probab.* **2**, 1024–1035.

Whittle, P. (1980), 'Multi-armed bandits and the Gittins index', *J. Roy. Statist. Soc.* **B42**, 143–149.

Whittle, P. (1988), 'Restless bandits: activity allocation in a changing world', *J. Appl. Probab.* **A25**, 287–398.

INITIAL DISTRIBUTION LIST

1. Research Office (Code 09).....1
   Naval Postgraduate School
   Monterey, CA 93943-5000

2. Dudley Knox Library (Code 013).....2
   Naval Postgraduate School
   Monterey, CA 93943-5002

3. Defense Technical Information Center.....2
   8725 John J. Kingman Rd., STE 0944
   Ft. Belvoir, VA 22060-6218

4. Richard Mastowski (Editorial Assistant).....2
   Department of Operations Research
   Naval Postgraduate School
   Monterey, CA 93943-5219

5. Distinguished Professor Donald P. Gaver 1
   Department of Operations Research
   Naval Postgraduate School
   Monterey, CA 93943-5219

6. Professor Patricia A. Jacobs.....2
   Department of Operations Research
   Naval Postgraduate School
   Monterey, CA 93943-5219

7. Kevin D. Glazebrook, Professor of Management Science 2
   Finance & Business Economics Group
   University of Edinburgh, School of Management,
   William Robertson Building, 50 George Square,
   Edinburgh, EH8 9JY, United Kingdom

8. Dr. Helen M. Mitchell.....2
   Department of Mathematics and Statistics
   University of Newcastle
   Newcastle upon Tyne NE1 7RU, United Kingdom

9. UK Technical Director,
   ONR Global (U. S. Office of Naval Research International Field Office).....1
   ATTN: Dr. James Greenberg
   ONRIFO, PSC 802, BOX 39, FPO-AE 09499-0039 USA

10. Professor Peter Denning ..... electronic copy
    Director, Cebrowski Institute
    Naval Postgraduate School
    pjd@nps.edu

11. Professor Phil Depoy ..... electronic copy
    Director, Meyers Institute
    Naval Postgraduate School
    pdepoy@nps.edu

12. Visiting Professor Greg Cox ..... electronic copy
    Operations Research Department
    Naval Postgraduate School
    gvcox@nps.edu

13. Professor Alan Washburn ..... electronic copy
    Operations Research Department
    Naval Postgraduate School
    awashburn@nps.edu

14. Professor Moshe Kress ..... electronic copy
    Operations Research Department
    Naval Postgraduate School
    mkress@nps.edu